ROUTING INFORMATION EXCHANGE

INVENTORS:

JOHN LEDDY

MICHAEL LLOYD

SEAN FINN

JAMES MCGUIRE

OMAR BALDONADO

HERB MADAN

PREPARED BY WILSON SONSINI GOODRICH & ROSATI

CROSS REFERENCE TO RELATED APPLICATIONS

This application claims priority to U.S. Provisional Application Nos. 60/241,450, filed October 17, 2000 and 60/275,206, filed March 12, 2001, which are hereby incorporated by reference in their entirety.

BACKGROUND OF THE INVENTION

Field of the Invention

This invention relates to the field of networking. In particular, the invention relates to measurements of network performance and optimization of network routing.

Description of the Related Art

The Internet today is comprised of groups of networks each run independently. Networks that build their own routing view of the Internet by connecting to multiple Internet Service Providers (ISPs) are Autonomous Systems (AS) and are assigned a unique 16 bit AS number. These networks exchange routing information about their Internet connectivity using BGP4.

As the Internet has grown, service providers have stratified into several categories. The trend has been moving toward 3 main categories: National/International Tier One ISP, Content Provider/Content Hoster and User Access Providers. This specialization has shown some weaknesses with the current system of distributed Internet decision-making.

National/International Tier One Internet Service Providers (ISPs)

As the Internet has evolved, there are fewer numbers of large Tier One providers able to compete by rebuilding their networks with newer high-speed routers, transmission gear, and access to dark fiber. As such, the difference between the top providers has decreased and the view of the Internet that is passed to multi-homed customers is very similar. Many routing decisions result from breaking ties or by enforcement of routing

policy by the customers, who attempt to balance load across available capacity. This results in decision making by end customers based on information that is becoming less and less differentiated and does not factor in the performance of the available paths.

The current state of Tier 1 ISPs evinces a need for an automated process for

5   generating routing decisions for such ISPs, based on up-to-date information on the performance of alternative paths.

Content Providers/Content Hosters

Content Providers need to multi-home in order to provide reliable high quality access to the Internet. As it is very difficult and expensive to connect directly with the

10   User Access Providers, connectivity to deliver content needs to be built by connecting to the large Tier One providers as well as by making localized policy decisions. As such, there is almost no coordination between Content Providers and their customers who get their connectivity through User Access Providers.

Content Providers also have the problem of delivering the majority of traffic from

15   their site to their user base. They need to make the decision of which of their directly connected ISPs can best deliver traffic for any particular user group. These decisions are usually made by allowing BGP to choose the "best route" (shortest number of network hops) and subsequently applying a local policy to hand tune the selections for particular destinations of interest. This, however, is not an automated process. Destinations with

20   large/important user groups are forced over paths that seem to provide better service proactively, and customers that complain are examined to see if a switch in paths could provide better service reactively. Issues of local outbound capacity with particular ISP connections often require traffic to be shifted off of even BGP "best route" paths, regardless of performance, in order to ease local congestion.

As such, there is a need for an automated process for selecting best routes for Content Providers to connect to particular user groups.

User Access Providers

User Access Providers need to manage inbound traffic to their user base from Content Providers. Today there are very poor mechanisms to control inbound traffic. Several mechanisms that are in use today are: Traffic Engineering (upgrading and purchasing new links from the "correct" ISPs); BGP padding to make a particular inbound path look bad to the entire Internet; and specific advertisements of small portions of a User Access Providers address space out different ISP links.

Each one of these methods has problematic operational issues. Traffic engineering requires an analysis at a particular time to select what ISPs should be used for future best cost and performance access to the Internet. Many times ISP service availability and local circuit installation require very long lead times. Service availability can often take 6 months or more, and contract terms are typically 1-3 years. By the time an ordered ISP service is available, the initial analysis may no longer be valid. Traffic flows may have changed and the performance of the ISP may have changed significantly.

BGP padding is a method of attempting to influence traffic flows in the Internet by artificially making a path look longer (more AS network hops). The User Access Provider will advertise its own connectivity to an ISP by appending its AS number multiple times in the BGP path. Networks making BGP "best route" decisions will see this path as lengthy and will be more likely to choose another available, shorter, path. This has the effect of reducing inbound traffic to the User Access Provider over the "padded" path. There are, however, several difficulties with this method. For instance, because the "padded" path is communicated to the entire Internet, there is no way to communicate a desired inbound

traffic policy to a particular traffic source. This causes significant amounts of traffic to be shifted away from the padded path (i.e., BGP padding allows very little granularity in how much traffic can be influenced to change paths). The results become even less granular as the number of Tier One ISPs decrease and become less differentiated. There is also no

5   way of communicating why the change is being requested and no way to take performance into account when a traffic source using BGP "best route" decisions receives a longer AS Path. As a normal operational procedure, a User Access Provider will "pad" a particular path and observe an initial shift in traffic. This initial traffic change may not be permanent. It may require several days as other networks and traffic sources adjust their

10   policies manually reacting to the change in traffic flows.

Another mechanism used by User Access Providers to influence inbound traffic is to use more specific IP route advertisements of their total address space. The Internet has incorporated CIDR (Classless Inter Domain Routing) into both its routing and forwarding decision-making. CIDR is a mechanism that allows multiple routes that are viable for a

15   particular destination to be present within the Internet. The path that is selected to forward traffic is the route with the more specific match of the destination IP address (longest match).

An example of this type of inbound policy is a User Access Provider that has 2 links, Link 1 and Link 2, each of which communicates with a different ISP, ISP1 and

20   ISP2, respectively. Ordinarily, the advertisements to ISP1 and ISP2 are identical. However, if there is more traffic than can be handled inbound on Link1 associated with ISP1 and available capacity on Link2 associated with ISP2, an inbound policy needs to be implemented to shift some traffic. Traffic engineering is not normally viable because of implementation times. Often a BGP Pad policy to artificially increase the network

25   distance associated with ISP1 will cause a significant amount of traffic to shift to ISP2 and

Link2. This may be more traffic than Link2 can carry and require the policy to be removed. To get finer granularity for the amount of traffic that is shifted, a more specific route advertisement is added to the ISP2 advertisement. This will cause traffic for a subset of the User Access Provider's customers to prefer ISP2 and Link2 inbound from the

5   Internet.

Although more control can be achieved over inbound traffic using this approach, it causes several problems. Management of the infrastructure is complicated since different groups of customers will have different performance and paths because of the fragmented policy. It also increases the global Internet route table size by requiring extra routes to be

10   carried by external networks to implement inbound policies. Many network infrastructures will ignore specific route advertisements that are "too small". Currently "too small" is an advertisement of a network route capable of addressing 4,096 hosts (/20 CIDR route). As such, this method will not provide fine granular control for providers with small amounts of address space. Additionally, as is the case with all the inbound

15   solutions, end to end performance is not able to be taken into account when shifting some flows from Link1 to Link2.

Looking Glass

In typical networks, in which routing paths are communicated between Autonomous Systems via BGP, the information about which outbound path has been

20   chosen from among the available paths is not communicated. Although this information is very useful for destination networks to know and act on, there is no mechanism or concept for the exchange of the resulting decisions. A troubleshooting tool that has been deployed by some networks to permit visibility into local routing decisions is called a Looking Glass (LG). The implementation of a LG is most often a WWW based user interface that

25   has a programmatic back end and can run a small number of queries on one of the

networks BGP routers. The deployment by networks of LG's is an example of the usefulness of the information. However, though an LG gives information about what path has been chosen to a particular destination by the network that deployed the LG, and the LG gives no information as to the performance or reason behind choosing a non BGP

5   "best route" path.

SUMMARY OF THE INVENTION

Some embodiments of the invention include network architectures and protocols to support enhancements to the decision making process of standard BGP. Some embodiments of the invention include a Routing Information exchange, or RIX. The RIX

10   comprises an overlay network which enables the exchange of routing information between Autonomous Systems (AS s) in an internetwork; one such example of an internetwork is the Internet. Embodiments of the RIX include one or more Points of Presence (POP's) distributed through the internetwork. These POPs may accept feeds from customer premise equipment. In some embodiments, these feeds may take the form of BGP4 feeds

15   which are supplied with local decisions made for forwarding traffic to the internetwork.

In some embodiments of the invention, the RIX may include a Path Selection eXchange (PSX), which allows decisions to be exchanged between autonomous systems about which internetwork paths have been selected for outbound traffic. In some embodiments of the invention, these decisions may take the form of one or more of the

20   following: default BGP selections, local "hand tuned" policies, or performance based decisions. In such embodiments, traditional BGP available path information may be enhanced with information about what paths have been chosen by other Autonomous Systems. Information supplied by the PSX may—by way of non-limiting example--be used for any one of the following: trouble shooting, traffic engineering, and enabling

policies. For instance, the information supplied by the PSX may be used to support

"symmetric routing", i.e., to keep the forward and reverse network paths equivalent.

Embodiments of the RIX include a Path Performance eXchange (PPX), which allows information about the measured performance of internetwork paths to be

5   exchanged within a localized area. In some embodiments, performance information may be sent to the PPX about internetwork destinations as measured over available paths. In some embodiments of the invention, the PPX may use this information from multiple sources to build a localized path performance database. In some such embodiments, this information may be encoded as a real time feed of performance data sent to customers in

10   the same localized area, or to users who are otherwise expected to experience similar performance. This information can be used to make local policy decisions incorporating performance.

Embodiments of the RIX include a Cooperative Routing eXchange (CRX), which enables additional policy information to be communicated between networks. Non-

15   limiting examples of such policy information include: information about why local policy decisions have been made; requests of policies from remote networks; performance information about particular paths; and informational status, all of which can be exchanged dynamically between networks.

In some embodiments, the components described above work together with

20   standard Internet Routers capable of BGP4, or with specialized equipment at customer premises, also referred to as Performance Aware Customer Premise Equipment (PACPE). However, as will be apparent to those skilled in the art, protocols other than BGP may be employed to send information between Autonomous Systems, PACPEs, and the RIX. By way of non-limiting example, these may be proprietary protocols or standard protocols,

25   such as IDRP (Inter Domain Routing Protocol). In some embodiments of the invention,

the RIX can also operate for networks supporting packet formats other than IPv4, for example IPv6 or OSI deployments. These and other embodiments are explained more fully below.

BRIEF DESCRIPTION OF THE FIGURES

Fig. 1 illustrates an architecture for the Routing Information Exchange, including an overlay network over multiple interconnected autonomous systems, according to some embodiments of the invention.

5   Fig 2 illustrates an example of a network configuration in which autonomous systems communicate routing performance information and policy decisions via the Routing Information Exchange according to some embodiments of the invention.

Fig 3. illustrates a BGP communities attribute as used to communicate information between autonomous systems and the Routing Information Exchange according to some 10   embodiments of the invention.

DETAILED DESCRIPTION

A. System Overview of the Routing Information eXchange

Some embodiments of the invention support a Routing Information eXchange, or RIX, comprising an overlay network 100, schematically illustrated in Figure 1, which is built to exchange routing information, performance information, decisions, and policy between groups of participating networks 102 104 106 in an internetwork. In an internetwork such as the Internet, these networks comprise Autonomous Systems 102 104 106. An AS 102 may be connected to other AS's 104 106 over paths 108 110 that may be physical or virtual. A BGP4 session may be run between the two AS's 102 104 to exchange a view of the Internet via that path 108. An AS 102 with connectivity to multiple AS's 104 106 takes these views of the Internet building a local combined view through local policies. This results in a decision for which outbound path to use for each possible destination.

The RIX 100 may be implemented in several embodiments using different protocols, which may be proprietary protocols or standard protocols, such as—by way of non-limiting example--IDRP (Inter Domain Routing Protocol). In some embodiments of the invention, the RIX 104 can also work for networks supporting packet formats other than IPv4, for example IPv6 or OSI deployments. Other protocols between networks which are compatible with the RIX 100 will be apparent to those skilled in the art.

In some embodiments of the invention, the RIX 100 may include a Path Selection eXchange (PSX), which allows decisions to be exchanged between autonomous systems 102 104 106 about which internetwork paths have been selected for outbound traffic. In some embodiments of the invention, these decisions may take the form of one or more of the following: default BGP selections, local "hand tuned" policies, or performance based

decisions--other decisions that may be exchanged between autonomous systems 102 104

106 will be apparent to those skilled in the art. In such embodiments, traditional BGP

available path information may be enhanced with information about what paths have been

chosen by other Autonomous Systems. Information supplied by the PSX may—by way of

5   non-limiting example--be used for any one of the following: trouble shooting, traffic

engineering, and enabling policies. For instance, the information supplied by the PSX

may be used to support "symmetric routing", i.e., to keep the forward and reverse network

paths equivalent.

Embodiments of the RIX 100 include a Path Performance eXchange (PPX), which

10   allows information about the measured performance of internetwork paths to be

exchanged within a localized area. In some embodiments, performance information may

be sent to the PPX about internetwork destinations as measured over available paths. In

some embodiments of the invention, the PPX may use this information from multiple

sources to build a localized path performance database. In some such embodiments, this

15   information may be encoded as a real time feed of performance data sent to customers in

the same localized area, or to users who are otherwise expected to experience similar

performance. This information can be used to make local policy decisions incorporating

performance.

Embodiments of the RIX 100 include a Cooperative Routing eXchange (CRX),

20   which enables additional policy information to be communicated between autonomous

systems 102 104 106. Non-limiting examples of such policy information include:

information about why local policy decisions have been made; requests of policies from

remote networks; performance information about particular paths; and informational

status.

The Internet currently functions by use of IPv4 as a network level addressing, formatting and forwarding protocol. Internet routing primarily relies on BGP4 as the standard network to network protocol for exchange of routing information. As such, the rest of this document focuses on using BGP4 as a underlying mechanism for the transport

5   and exchange of information necessary to implement the concept of the RIX (PPX, PSX and CRX) for IPv4 networks within the current Internet or Intranet's. However, the present invention is not limited to BGP4 as the sole gateway protocol between Autonomous Systems 102 104 106, and other alternatives will be apparent to those skilled in the art.

10   B. IPv4 and BGP4 RIX Implementation

In some embodiments of the invention, the RIX 100 includes one or more Points of Presence (POPs) deployed within the Internet. In some embodiments, these POPs have the capability of accepting BGP4 connections from customer equipment 112 114 116, which may be routers or PACPE (Performance Aware Customer Premise Equipment). A

15   given Autonomous System may include sub-networks for many different organizations, each of which may have one or more PACPEs. PACPEs are further described in U.S. Provisional Application Nos. 60/241,450, filed October 17, 2000 and 60/275,206, filed March 12, 2001, all of which are hereby incorporated by reference in their entirety. In some embodiments of the invention, the BGP4 feed sent to the RIX 100 by the customer

20   112 114 116 is a standard BGP4 feed which includes the result of the local decisions made for forwarding traffic to the Internet. This is the same feed a customer would establish with a network selling Internet access (BGP4 Transit Feed). This feed establishes a base level of communication between the customer and the RIX 100. It also establishes

information used to build the PSX. The information from multiple customer feeds is parsed by the RIX 100 into information specific to each customer.

As an illustrative, non-limiting example, consider the network illustrated in Figure 2. A network 192.100.10.X (AS 1) 200 has two ISPs (AS10 and AS20) 202 204 a second network 192.200.20.X (AS2) 206 has two ISPs (AS20 and AS30) 204 208, a third network 192.300.30.X (AS3) 210 has two ISPs (AS 40 and AS50) 212 214. Each network has a eBGP4 connection to the RIX 100 sending their information. The following information is sent from the networks 200 206 210 to the RIX 100:

Network1 to RIX (AS1):

192.100.10.X	AS Path:	AS1
192.200.20.X	AS Path:	AS10, AS30, AS2
192.300.30.X	AS Path:	AS10, ASX, ASY, AS40, AS3
Destination A	AS Path:	AS20...
Destination B	AS Path:	AS10...
Destination C	AS Path:	AS20...

Network2 to RIX (AS2):

192.100.10.X	AS Path:	AS20 AS1
192.200.20.X	AS Path:	AS2
192.300.30.X	AS Path:	AS20, ASZ, AS50, AS3
Destination A	AS Path:	AS30...
Destination B	AS Path:	AS30...
Destination C	AS Path:	AS20...

Network3 to RIX (AS3):

192.100.10.X AS Path: AS50, ASZ, AS20, AS1

192.200.20.X AS Path: AS50, ASZ, AS20, AS2

192.300.30.X AS Path: AS3

Destination A AS Path: AS40...

5   Destination B AS Path: AS40...

Destination C AS Path: AS50...

The RIX 100 then stores Network Specific Information which may include one or

more of the following Reverse Path Information:

10

Network1 Reverse Path Information

192.100.10.X AS Path: AS2, AS20 AS1   Network2 Path

192.100.10.X AS Path: AS3, AS50, ASZ, AS20, AS1   Network3 Path

15   Network2 Reverse Path Information

192.200.20.X AS Path: AS1, AS10, AS30, AS2   Network1 Path

192.200.20.X AS Path: AS3, AS50, ASZ, AS20, AS2   Network3 Path

Network3 Reverse Path Information

20   192.300.30.X AS Path: AS1, AS10, ASX, ASY, AS40, AS3 Network1 Path

192.300.30.X AS Path: AS2, AS20, ASZ, AS50, AS3   Network2 Path

C. RIX BGP4 Annotations

In some embodiments, the device on the customer premises 112 114 116 establishing a BGP4 session with the RIX 100 is a PACPE, thus enabling additional interactions with the RIX 100. In some such embodiments, the base BGP4 session has

5 new information added to communicate data flows that enable RIX 100 components and enhance the operation of a PACPE 112 114 116. In some embodiments of the invention, this information is carried in the BGP Communities attribute in the feed to and from the RIX 100. A BGP Community 300, as illustrated in Figure 3, is a 32 bit quantity: the first 16 bits 302 comprise an AS number and the second 16 bits 304 comprise a value whose

10 interpretation is defined within that AS (AS: value). A private AS is a reserved set of AS numbers that can be used privately; the reserved set includes values from 64512 to 65535.

D. Equivalence Class Feed (from RIX to PACPE)

Some embodiments of the invention include an Equivalence Class (EC) feed, providing a PACPE 112 114 116 additional information about the structure of destination

15 networks. Equivalence Classes comprise clusters of network prefixes which are grouped together. In some embodiments of the invention, network prefixes are grouped into an equivalence to reflect similar performance characteristics. Equivalence Classes are further described in U.S. Provisional Application No. 60/241,450, filed October 17, 2000, and U.S. Provisional Application No. 60/275,206, filed March 12, 2001, all of which are

20 hereby incorporated by reference in their entirety.

The implementation of BGP4 within the Internet has been successful reducing the rate of growth in the number of routes a router needs to carry; networks today are encouraged to advertise the largest possible aggregation of network routes (smallest number of routes) when exchanging information with other networks. However, this

causes information about geographic deployment and connectivity of smaller aggregations to be lost to the general Internet. The EC feed is recognition that current Internet priorities, such as to reduce route table size, are in opposition to selecting the best performance route to specific destinations.

5   In some embodiments of the invention, the EC feed to PACPE 112 114 116 comprise advertisements of destinations that have performance paths and should be treated as a unit for making measurement decisions. They can be more specific than a BGP advertisement, fragmenting the Internet BGP4 advertisement into smaller blocks with independent performance, or they may comprise multiple independent advertisements that

10   are tagged as a performance group. Cases in which an EC Tag may be communicated to PACPE 112 114 116 are described below:

Prefix is Unique to the EC feed

A route is built with an EC destination and network mask. It is given an Origin AS

15   associated with the RIX 100 and advertised into the eBGP feed from the RIX 100 to the customer. In some embodiments of the invention, the route is tagged with a community string using a private AS (AS 65001) and a value 0. If the EC is associated with other EC's in a performance group, a second community may be added to the string with the same private AS (AS 65001) and a value that is unique to all other EC's in the group. Any

20   information other than the Community values can be rewritten by other data flows.

EC Network, Mask (Stand Alone EC)

AS Path: Origin AS RIX(65534)

Community String: (AS 65001:0)

25

EC Network, Mask (EC part of group ID)

AS Path: Origin AS RIX(65534)

Community String: (AS 65001:0), (AS 65001:group ID)

5

Prefix Exists as Part of Another Feed

In some embodiments of the invention, if there is already a route advertisement to the PACPE 112 114 116 from the RIX 100 from another data flow, to make that destination an EC, the route may be tagged with a community string using a private AS (AS 65001) and a value 0. If the EC is associated with other EC's in a performance group, a second community is added to the string with the same private AS (AS 65001) and a value that is unique to all other EC's in the group. If the original route advertisement is deleted, a new route is created and advertised as when the "Prefix is Unique to the EC Feed" as described above.

15

Destination Network, Mask (Stand Alone EC)

AS Path: Original AS Path

Community String: (AS 65001:0)

20   Destination Network, Mask (EC part of group ID)

AS Path: Original AS Path

Community String: (AS 65001:0), (AS 65001:group ID)

E. Performance Measurements from PACPE to the RIX

In some embodiments of the invention, data may be sent to a RIX 100 from a

PACPE box 112 114 116 on a customer premise that is measuring performance.

Performance measurements across available ISP paths are encoded and sent to the RIX

5   100 as community values and associated with the active route.

In some embodiments of the invention, the performance value is sent as (FH AS:

value). The "First hop (FH) AS" is the first ISP's AS over the measurement path. The

value is an encoded measure of performance and defined as a <type, argument> value pair.

Measurements Associated with a BGP Selected Route

10   In some embodiments of the invention, if the measurement values are associated

with a route currently being advertised to the RIX 100, the route is tagged with the BGP

communities containing the performance data. Performance data may be inserted as a new

route advertisement with the new Community values. If the route is changed, the

measurement values do not need to be moved to the new route advertisement. A new set

15   of periodic performance data can be sent to the RIX 100 when available in the new

advertisement.

Destination Network, Mask

20   AS Path: Original AS Path

Community String: (FH AS1: value1),... (FH ASx: valuex), Original Community

String

Measurements Associated with a Performance Selected Route

If the measurement values are associated with a route that has been chosen based

25   on a local performance decision, the original routing information may not be available to

the PACPE 112 114 116. If the path information is available the case should be treated as in the "Measurements Associated with a BGP Selected Route" scenario described above. If the information is not available, a route advertisement may be made for the Performance Selected Route. If the route is changed or a new route added, the measurement values do

5   not need to be moved to the new route advertisement. A new set of periodic performance data can be sent to the RIX 100 when available in the new advertisement.

Destination Network, Mask

10   AS Path: Original FH AS, RIX(65534), origin AS

Community String: (FH AS1: value1),... (FH ASx: valuex), Original Community String

Measurements Associated with EC's not in BGP:

If the measurement values are associated with an EC but the PACPE 112 114 116

15   has not installed the EC as a route into the customer forwarding routers, the performance data may be sent to the RIX 100 In some embodiments of the invention by building a route advertisement. The advertisement is for the EC using the AS of the RIX as the origin AS and inserting the BGP communities with the performance measurements associated with each FH path.

20
Destination Network, Mask

AS Path: Origin AS RIX(65534)

Community String: (65001:0), (65001:ID), (FH AS1: value1),... (FH ASx: valuex)

25

F. Performance Measurements from RIX to PACPE

In some embodiments of the invention, the RIX 100 takes performance data received from PACPE feeds to the RIX 100 and aggregates measurements into a value that is representative of ISP performance in a localized area. This information is then relayed
5   to PACPE 112 114 116. PACPE devices 112 114 116 use the information to make decisions about performance-based First Hop ISP outbound path choices.

Another type of performance data that may be sent to the PACPE 112 is the value of the performance advertisement to the RIX 100 from another customer PACPE 116. This is the value associated with a local PACPE measurement advertised to the RIX 100
10   from the customer that owns the prefix to the receiver of the feed. This is a form of Cooperative Routing that relays information between source and destination networks. The received value for performance from a PACPE may be used to encode a value as a BGP Community in the form (65100: value). In some embodiments of the invention, the community is then tagged onto routes in any PACPE feeds that contain the prefix owned
15   by the original advertiser.

Prefix is Unique to the EC feed:

If the prefix is unique to the EC feed, performance information can be tagged by adding BGP community values. In some embodiments of the invention, the performance
20   value is sent as (FH AS: value). The "First hop (FH) AS" is the first ISP's AS over the measurement path. The value is an encoded measure of performance and defined as a <type, argument> value pair.

Destination Network, Mask (Stand Alone EC)

25   AS Path: Origin AS RIX(65534)

Community String: (AS 65001:0), (FH AS1: value1),... (FH ASx: valuex)

Destination Network, Mask (EC part of group ID)

AS Path: Origin AS RIX(65534)

5   Community String: (AS 65001:0), (AS 65001:group ID), (FH AS1: value1),... (FH

ASx: valuex)

Prefix Exists as Part of Another Feed

If the measurement values are associated with a route currently being advertised to

10   the PACPE 112 114 116, the route may be tagged with the BGP communities containing

the performance data. Performance data can be inserted as a new route advertisement with

the new BGP Community values. If the route is changed, the measurement values do not

need to be moved to the new route advertisement, and a current set of performance data

can be sent from the RIX 100.

15

Destination Network, Mask

AS Path: Original AS Path

Community String: (FH AS1: value1),... (FH ASx: valuex), (65100: value),

20   Original Community String

G. Cooperative Data feed from PACPE to RIX

Some embodiments of the invention support Cooperative Routing between source

and destination networks. The Cooperative Routing function in the RIX 100 enables the

encoding of information that is communicated intact about network pairs (destination

25   prefixes to source networks). This information can be used to communicate information

including, but not limited to any one of the following: hints, policy, performance, requests, status, and information. Each of these Cooperative Routing verbs relate information about the tagged route.

In some embodiments of the invention, The Cooperative Routing data verbs are

5   carried as BGP Community of the form:

(Cooperative Private AS: value). The values are defined for each Cooperative Private AS and can take the form of <type, value>.

In some embodiments of the invention, Cooperative Routing Information may be defined

10   as below. This is provided as an example, as many other suitable permutations of Private AS values will be apparent to those skilled in the art:

(65001-65100: value)

15   65001 EC

65002 Requesting Symmetric AS Path Routing: 0

65003 Prefer paths with this AS (1st priority): AS

65004 Prefer paths with this AS (2nd priority): AS

65005 Prefer paths with this AS (3rd priority): AS

20   65006 Avoid paths with this AS (1st priority): AS

65007 Avoid paths with this AS (2nd priority): AS

65008 Avoid paths with this AS (3rd priority): AS

65009 DOS attack (Black Hole): 0

65010 DOS attack (Rate Limit): 0

25   65011 DOS attack (Informational): 0

65012 Packet Loss Unacceptable: value (encoded Packet Loss number)

65013 Jitter Unacceptable: value (encoded Jitter Number)

65014 Scheduled Outage: value (date,time,duration)

65015 – 65099 Reserved

65100 Performance data:value

5   H. Cooperative Data Feed from RIX to PACPE

In some embodiments of the invention, the Cooperative Routing information that is received by the RIX 100 is parsed to aggregate information from all customers about a specific customer. This information can then be used to annotate the BGP4 feed to the PACPE 112 114 116. For example, a customer that wishes to inform a remote network

10   that they recommend preferring Internet paths in the reverse direction that contain a particular transit ISP (AS10) may take the route advertisement to the remote network and insert a BGP Community value of (65003:AS10). The RIX 100 takes this information and communicates it to the POP where the remote network has a RIX feed. The (65003:AS10) BGP community can then be assigned to the route in the remote networks feed associated

15   with the customer network. A PACPE 112 114 116 receiving this information can make a local decision about how much weight to put on the request, from ignoring it to following absolutely.

In an illustrative example, Network1 sends its BGP4 feed to the RIX with the

20   routes to Network2 assigned a Community value of (65003:AS10). Network1 is informing Network2 that it prefers reverse paths that contain AS10.

Network1 to RIX (AS1):

192.100.10.X AS Path: AS1

25   192.200.20.X AS Path: AS10, AS30, AS2

Community String: (65003:AS10)

Network2 to RIX (AS2):

192.200.20.X AS Path: AS2

192.100.10.X AS Path: AS20, AS1

Community String:

5

BGP feed from RIX to Network1 and Network2

RIX to Network1 (AS1)

192.200.20.X AS Path: AS2, AS20, AS1

10   Community String:

RIX to Network2 (AS2)

192.100.10.X AS Path: AS1, AS10, AS30, AS2

Community String: (65003: AS10)

15

I.   Conclusion

        The foregoing description of various embodiments of the invention has been

presented for purposes of illustration and description. It is not intended to limit the

invention to the precise forms disclosed. Many modifications and equivalent

20   arrangements will be apparent.